What are you hiding? Initial validation of the reaction time-based searching concealed information test

Koller, Dave ; Hofer, Franziska ; Grolig, Tuule ; Ghelfi, Signe ; Verschuere, Bruno

Abstract: The reaction time-based concealed information test (RT-CIT) has been used to judge the veracity of an examinees claim to be naïve by using RTs to test for recognition of relevant details. Here, we explore the validity of the RT-CIT to generate new knowledge about the incident—the searching CIT. In a mock terrorism study (n = 60) the RT-CIT not only allowed to link suspects to known crime details, but also allowed to reveal new crime details well above chance. A simulation study confirms the potential of the searching RT-CIT and identifies conditions under which it performs best. We used an archival dataset that met these conditions (high CIT effect, large number of item repetitions), and found better item classification performance than in the mock terrorism study. The searching RT-CIT could be a new, promising investigative tool to reveal new (e.g., crime) details to the investigative party.

DOI: https://doi.org/10.1002/acp.3717

Originally published at:
Koller, Dave; Hofer, Franziska; Grolig, Tuule; Ghelfi, Signe; Verschuere, Bruno (2020). What are you hiding? Initial validation of the reaction time-based searching concealed information test. Applied Cognitive Psychology, 34(6):1406-1418.
DOI: https://doi.org/10.1002/acp.3717

Received: 9 January 2020 | Revised: 19 June 2020 | Accepted: 23 June 2020

DOI: 10.1002/acp.3717

WILEY

RESEARCH ARTICLE

What are you hiding? Initial validation of the reaction time-based searching concealed information test

Dave Koller[1,2] | Franziska Hofer[3,4] | Tuule Grolig[1] | Signe Ghelfi[3] | Bruno Verschuere[2]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland

[2]Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

[3]Zurich State Police, Airport Division, Research and Development, Zurich, Switzerland

[4]Brainability, Developing Human & Organizational Potentials, Zurich, Switzerland

Correspondence
Dave Koller, Department of Psychology, University of Zurich, Binzmühlestrasse 14/Box 22, 8050 Zurich, Switzerland.
Email: d.koller@psychologie.uzh.ch

Funding information
Federal Office of Civil Aviation; Zurich State Police

Summary

The reaction time-based concealed information test (RT-CIT) has been used to judge the veracity of an examinees claim to be naïve by using RTs to test for recognition of relevant details. Here, we explore the validity of the RT-CIT to generate new knowledge about the incident—the searching CIT. In a mock terrorism study (n = 60) the RT-CIT not only allowed to link suspects to known crime details, but also allowed to reveal new crime details well above chance. A simulation study confirms the potential of the searching RT-CIT and identifies conditions under which it performs best. We used an archival dataset that met these conditions (high CIT effect, large number of item repetitions), and found better item classification performance than in the mock terrorism study. The searching RT-CIT could be a new, promising investigative tool to reveal new (e.g., crime) details to the investigative party.

KEYWORDS

application, deception, external validity, memory detection, searching concealed information test (CIT)

1 | INTRODUCTION

By testing a suspect on crime information that only a perpetrator, a witness or a victim could have, the concealed information test (CIT) also known as guilty knowledge test (Lykken, 1959) can connect an examinee to knowledge about the crime.

To illustrate how the CIT can be used, imagine the following scenario: Two burglars broke into a storage hall of Pravay (a chemical plant) with the use of a crowbar. They stole large quantities of concentrated sulfuric acid that can be utilized to synthesize explosives. Based on low quality closed-circuit television footage and a terror watch list, the police bring in a suspect for questioning. He denies all knowledge about the break in. With the information the police officers have about the crime, they can construct a *known solution* CIT by taking the true crime information (Pravay, crowbar, sulfuric acid; so-called *probes*) and adding plausible alternatives (company names, other tools often used to break in, different chemicals; so-called *irrelevants*). When asked about the crime, a naïve person cannot distinguish between the probes and the irrelevants and therefore does not show a systematic difference regarding the response to the stimuli. On the other hand, a knowledgeable person shows recognition of the probes and may attempt to hide that (Klein Selle, Verschuere, Kindt, Meijer, & Ben-Shakhar, 2017). These processes are typically accompanied by an increase in skin conductance response (SCR), response times, and P300 amplitude as well as a decrease of the heart rate and respiration line length; all of which can be used to classify individuals into knowledgeable/naïve well above chance (e.g., Meijer, Klein Selle, Elber, & Ben-Shakhar, 2014; Seymour, Seifert, Shafto, & Mosmann, 2000; Suchotzki, Verschuere, van Bockstaele, Ben-Shakhar, & Crombez, 2017). Classification performances range from an area under the curve (AUC) of AUC = 0.74 for heart rate to AUC = 0.88 for P300 amplitude (Meijer et al., 2014) with response times achieving

AUC = 0.82 (Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakhar, 2016).[1]

Detecting if a suspect is involved in the crime of interest is often not enough. In a real-life scenario similar to the described break in, the police are not just interested in assessing whether the suspect may be involved in the burglary, but they are also, perhaps primarily, eager to prevent the attack. For that purpose, it would be helpful to get an answer to such questions as: Who is the second burglar? Are there more people involved? Where is the explosive synthesized and stored? Where and when do they intend to execute the attack? The police might have a list of critical infrastructure and possible targets with many casualties, but to act effectively with limited time and resources, the police needs to know the target of the upcoming attack. The approach to tackle this challenge is called the *searching CIT* as the police is searching for the probe amongst a set of probable alternatives. Contrary to the known solution CIT, the police do not know the crime information in the searching CIT. However, for the searching CIT to work, it is crucial to have a set of items that includes the true crime information with a very high probability. If the actual crime information is included in the CIT, a knowledgeable person does still recognize this information. The person tries to hide the knowledge which leads to the aforementioned effects (e.g., increased RTs). Based on the observed data, the searching CIT tries to classify each item as either being crime irrelevant or crime relevant (i.e., an irrelevant item or a probe item). The difference to the known solution CIT arises in the way the data are analyzed.

The idea of the searching CIT is not new. Autonomic measures have been used to extract information from groups of participants with shared complete (e.g., Breska, Zaidenberg, Gronau, & Ben-Shakhar, 2014; Meijer, Bente, Ben-Shakhar, & Schumacher, 2013) or partial crime knowledge (e.g., Elaad, 2016) in experiments, and Japanese law enforcements regularly use autonomic measures based searching CIT (Osugi, 2011).

One of the few single-subject searching CIT studies was conducted by Meixner and Rosenfeld (2011) using EEG. Participants were assigned either to the guilty condition in which they were asked to plan a terror attack with three testable crime information (the type of attack: bomb; location: Houston; time: July) or the innocent condition in which they planned a vacation. These items were tested against five irrelevant items in each information category. Therefore, the probes were always present among the tested items. In order to find the probes, for each participant, they compared the two items with the largest mean P300 amplitude within each category. If the difference (measured by comparing bootstrapped means) between these items was sufficiently big, it was concluded that the item with the largest mean P300 is crime relevant (probe), otherwise it was concluded that there is no probe item in this category and participant. This algorithm achieved a probe classification accuracy of 67% (chance performance was 20%). The technical requirements of EEG-systems and the trained personal needed to administer an EEG are barriers to applying it in practice and limiting factors when it comes to scalability. The physiological CIT, although cheaper, cannot be scaled up easily for the same reasons. The RT-CIT, however, requiring only a computer and data collection and analysis possibly being fully automated, allows for remote and parallel testing with little additional resources needed.

As far as we know, the present study is the first to explore the validity and applicability of the reaction time-based searching CIT (searching RT-CIT). We evaluate two searching algorithms for the scenario where the investigators themselves do not know what the critical details are (i.e., searching RT-CIT). In contrast to the common known-solution CIT in which the examinee is tested for critical details that the examiner knows are related to the crime, we pretend to be ignorant about the crime and aim to classify the items as crime relevant/irrelevant and use this to classify participants as guilty/innocent in an airport setting using a mock crime paradigm (Study 1). Based on these results, the performance of the algorithms under different conditions was explored using a simulation study (Study 2). Finally, the algorithms were cross validated on independent data (Study 3).

The two algorithms we evaluate are inspired by Meixner and Rosenfeld (2011), and Noordraven and Verschuere (2013). We expect above-chance classification performance for the items (Hypothesis 1a) and in a second step for participants (Hypothesis 1b), based on the item classification for both algorithms.

2 | STUDY 1: APPLYING THE SEARCHING RT-CIT IN AN AIRPORT SETTING

Study 1 used a mock crime paradigm at an international airport, with a guilty group that planned a mock terror attack and partially executed it, and an innocent control group (see Procedure section). Two searching RT-CIT algorithms were used to detect crime relevant information and to classify participants. A priori, we expected both algorithms to show above chance classification performance for items and participants, but we had no predictions when it came to comparing the two algorithms.

To draw conclusions about the searching RT-CIT in an airport setting, we first need to validate the known solution RT-CIT in that setting; an environment with high security standards (enforced by the police) that, in addition, is relatively unfamiliar to participants and therefore likely to cause higher agitation levels in all participants than a laboratory setting at a university does. Thus, we predict a larger standardized probe-irrelevant difference in RTs ($\frac{M(\text{probe})-M(\text{irrelevant})}{SD(\text{irrelevant})}$) as introduced by Noordraven and Verschuere (2013) and here forth called *CIT-effect*) for participants in the guilty vs. innocent group (Hypothesis 2a). In a similar vein, we expect the guilty and innocent classification accuracy based on the CIT-effect to be greater than 50% (Hypothesis 2b).

As a secondary aim, Study 1 also investigated potential effects of richer memory traces of past actions compared to intentions (e.g., Cohen, 1981) on the CIT-effect. Although it has been shown that reaction times can be used to detect intentions with the CIT (Noordraven & Verschuere, 2013), it is unknown if there is a difference in how well past actions and intentions can be detected using RTs. The insight we gain is of high practical relevance as it will show if the RT-CIT is suitable for exposing planned criminal actions before the crime is committed which is especially important in the context of terrorism.

2.1 | Method

The experiment was approved by the ethical committee of the Faculty of Arts and Social Sciences of the University of Zurich (Approval number: 2018.2.11). The study is exploratory,[2] data and code can be found on osf.io/69yrj.

2.2 | Participants

Participants were 60 students from the University of Zurich (*M* age = 22.5 years; *SD* = 3.1 years, range 19–32 years, 47 female). To end up with a balanced design, we recruited until we had 60 participants after applying the preregistered exclusion criteria. Of all the tested participants (*n* = 68), 8 were excluded (1 due to poor performance in the task [more than 50% errors in at least one item category], 6 exceeded the two-error-limit in the post-CIT recognition task, and 1 participant failed both criteria). Participants were recruited via participants' mailing-list, postings on bulletin boards at the university and advertisements in lectures. The participants were enrolled to the study when the following inclusion criteria were met: age between 18 and 35, high school degree or higher, and fluent in German. Before the experiment started, all participants were asked to read and sign the informed consent. It was clearly stated that the participation is voluntary, and withdrawal is possible at any time during the course of experiment with full compensation. All participants received 20 CHF (≈20.40 USD) or course credits for 1.5 hr of study participation (participants' choice). All participants were told that they will earn an additional 5 CHF (≈5.10 USD) if they can complete their task without being accused of anything (indicated by a search of their hand luggage). They were specifically instructed that simply being suspected is not enough to lose this bonus. However, independent of their performance, all participants received the additional 5 CHF.

Half of the participants were asked to plan for a mock-crime (guilty suspects; *n* = 30; *M* age = 22.10 [*SD* = 2.80]; 24 female; 23 right-handed). The other half was asked to plan for a non-criminal act (innocent suspects; *n* = 30; *M* age = 22.83 [*SD* = 3.34]; 23 female; 27 right-handed).

2.3 | Procedure

The experimental procedure consisted of four phases—*planning phase*, *execution and interception*, *RT-CIT*, and *target and probe recognition*.

2.3.1 | Planning phase

All participants were contacted by e-mail and requested to bring a self-packed cabin bag for a day trip with an airplane. It was explicitly stated that the bag must not contain any forbidden items. The participants were informed that new security measures and communication protocols between different divisions of the airport police had been introduced. Participants were told that this study is part of an airport security check to test these measures. Upon arrival to the airport, an experiment leader welcomed the participant and brought him/her to an office room where he/she read the instructions on a sheet of paper. The participants in the guilty group were told to take part in a mock-terror attack to test the newly implemented security measures. The innocent group was given no additional information beside that they are to test the efficacy of the new security protocol. All participants received mock flight documents issued to their name and a map of the airport. They were given up to 7 min to plan their tasks.

The participants in the guilty condition were instructed to go to a location marked on the map where they will find an envelope with a code word on it. The envelope contained a numbered key to a safe deposit box at the airport in about 5 min walking distance. In the safe deposit box, they found two items which they should take within their hand luggage. Item 1 was to be smuggled through the security check and handed over to a confederate (airside), whereas Item 2 was to be used as a sign to be recognized by the confederate. The purpose of each item was explained in the instructions (see Appendix A). The guilty participants were instructed to carefully hide the items in their luggage in order to avoid detection. After crossing the security check, they were asked to proceed to a specific gate, take out Item 2, and wear it or hold it visibly until they are approached by the confederate which they will recognize by a certain feature (e.g., red shoes) and hand over Item 1. The guilty participants were also instructed that in case nobody approached them within 5 min, they should deposit the item at a certain place and then meet the experiment leader at a location described in the instructions (see Table 1 for all the items). To control for possible

TABLE 1 Items used in the RT-CIT sorted by enactment (enacted task/intention) and item category

Enactment	Item category	Item
Enacted	Code word	Gondola, magnet, board, anchor, stamp, clip
Enacted	Number of safe deposit box	Six unique random numbers from 1 to 98 (newly generated for each participant)
Enacted	Item 1, to be smuggled	USB-stick, watch, liquid, phone, powder, wallet
Enacted	Item 2, used for recognition	Thermos, map, book, headphones, pen, bag
Intended	Gate	A81, B34, C06, D55, E27, F19
Intended	Confederate feature	Red shoes, grey pants, brown jacket, blue scarf, white hat, golden sunglasses
Intended	Item placement	In trash can, in PET container, in aluminium recycling bin, under seat, behind pillar, behind door
Intended	Meeting spot	Information desk, Hour Passion, Center Bar, Amavita, Marché, NZZ Café

item effects, the information used in the instructions was balanced over the course of the experiment in the sense that within each category, every item was used exactly five times.

The innocent participants were instructed to go to four specific stores at the airport that were indicated by numbers on a map they received. In each store, they were to select an item and write down its name and price on a sheet of paper received from the experiment leader. After completing their task, the participants were asked to use the provided flight documents and go through the security check. Thereafter, they were told to have 15 min waiting time to be filled as they wanted. Next, the participants were requested to meet the experiment leader at the baggage check. This procedure ensured that innocent participants had to complete a task that took a similar amount of time as the mock crime, that lead them to the same areas of the airport, and that also included orienting and planning based on a map and navigating through the airport preferably relying on their memory.

All participants (innocent and guilty) executed their tasks as planned up until they were about to use their flight documents (enacted tasks). The activities that should have followed after the security check were considered planned future acts (intentions).

2.3.2 | Execution and interception

Immediately before participants used their flight documents, they were intercepted by a confederate posing as an undercover police officer and brought to an office room where they were asked to complete a security test (RT-CIT). The items the guilty participants encountered before this interception (code word, number on the key, item 1 and item 2) were considered to be part of the enacted tasks whereas gate, recognition feature, item placement, and meeting spot were considered intention-items.

2.3.3 | RT-CIT

Upon entering the CIT-room, participants were orally instructed to turn off/mute their phone and to put it aside together with their bag (and possibly other items). They were asked to sit down in front of the computer and the confederate posing as a police officer informed them that the following test was designed to determine if he/she has knowledge about a crime. They were told that all the instructions will be presented on the screen but if they had a question at any point, he/she could ask the police officer who stayed in the same room in hearing distance but not visible.

The RT-CIT was programmed with MATLAB version 9.4.0 (The MathWorks, 2018) with the Psychtoolbox extension version 3.0.14 (Brainard, 1997) on a Dell Latitude E6530 with a 15.6″ screen running on Windows 7 (Service Pack 1). Participants were seated approximately 50 cm from the screen.

The RT-CIT consisted of eight item categories (4 past action categories, 4 intention categories), with six items per category (1 probe, 1 target, 4 irrelevant). In the beginning, participants were asked to learn the target items to which they should respond with "YES, this is connected to the crime" in the subsequent task. Participants could end the learning phase on their own as soon as they felt well prepared. They were then presented with all 48 items of the test and were asked to click on the target items with the mouse to ensure that the items have been memorized. The selected items were highlighted to help the participants to keep track of their choices. Participants with more than one error in this recognition test received feedback that they have made more than one error and that they should learn the target items again. This procedure was repeated until no more than one error was made.

After passing the target recognition test, the reaction time trials started. On each trial, a single item was presented in the middle of the screen, either a probe, a target or an irrelevant item. Participants were asked to indicate as fast and as accurately as possible whether the item has a connection to the crime by pressing either "e" or "i" on the keyboard. The answers were assigned in a way that NO was always pressed with the participants' dominant hand. The response–stimulus-interval varied randomly from 500 to 1,000 ms.

There were three practice blocks of 32 trials each to familiarize the participants with the task before the actual test trials started. Every item was presented twice in this practice phase. The first practice block had a response time limit of 10 s. The items remained on the screen until the participant pressed a button or the time limit was reached. For the second and third practice phases, participants were instructed to respond within 1.2 s and 0.8 s respectively. A red "TOO SLOW" message appeared above the stimulus if the participants did not respond within those instructed time intervals. However, every response given up until 1.5 s after the stimulus presentation was recorded. Participants also received feedback on whether they responded correctly (i.e., YES to targets, NO to probes and irrelevants) by indicating errors with the display of a red "X" below the stimulus. The actual test phase consisted of 20 blocks. In each block, every item is presented once, resulting in 960 trials in total which took approximately 23 min to complete. Participants were given the opportunity for two short, self-paced breaks after blocks eight and fifteen.

2.3.4 | Target and probe recognition

After the RT-CIT ended, there was another target recognition test to ensure that participants did not forget the targets during the test. Additionally, participants in the guilty group had to complete a probe recognition test. They were told that only the guilty participants see this test, that it is needed to evaluate the study properly, and that they should, therefore, answer truthfully.[3]

2.4 | Searching CIT algorithms

In contrast to the known solution CIT, we assume to be ignorant of the guilt of the participant and about the items that were involved in

the crime. We applied two searching-CIT algorithms aimed to classify each item as relevant/irrelevant and each participant as guilty/innocent.

2.4.1 | Standardization algorithm

Conceptually, the first algorithm that we tried is one in which every item is treated as the possible probe and compared to all the other items in its category. Items with CIT-scores above a certain cut-off are classified as probes, the others as irrelevants. Based on the idea of standardizing the probe-irrelevant difference within a participant (Noordraven & Verschuere, 2013), this algorithm uses within category standardized CIT-scores. For each participant and every item *i* in item category *j*, the CIT-scores are calculated as $dCIT_{i,j} = (M(RT_{i,j}) - M(RT_{k \neq i,j}))/SD(RT_{k \neq i,j})$. For participant-classification, we use the mean of the largest $dCIT_{ij}$ scores of each category $dCIT_p = M(\max(dCIT_{i,j}))$ and classify a participant as "guilty" if $dCIT_p$ is above a certain threshold and as "innocent" otherwise.

2.4.2 | First to second bootstrap algorithm

The second algorithm we applied on our RT data has been successfully used in the P300 CIT (Meixner & Rosenfeld, 2011). The rationale behind this algorithm is that the probes should have the largest RT and that they should only be classified as probe if the difference to the item with the second largest RT is sufficiently big. In a first step, the item with the largest mean RT in each category is identified for every participant. These items are considered possible probes. The item with the second largest RT is presumed irrelevant and will be used for comparison. All other items are also considered irrelevants, but they are ignored for the rest of the algorithm.

In a second step, 2000 bootstrap sample means are calculated for the possible probe and the presumed irrelevant for each category. A sample is created by drawing (with replacement) as many RTs from the responses to a given item as there are valid trials for that item. In each of these 2000 iterations, the mean RT for the possible probe item is compared to the mean RT of the irrelevant item. The *m*-th bootstrap sample of category *j* is denoted $Pboot_{j,m}$ and $Iboot_{j,m}$ for the possible probe and the presumed irrelevant respectively. The possible probe of category *j* is then classified by $boot_j$, the percentage of iterations in which its mean RT was larger than the mean RT for the irrelevant item ($boot_j = \frac{count_m\left(M\left(Pboot_{j,m}\right) > M\left(Iboot_{j,m}\right)\right)}{2000}$). Participants are classified based on the mean of those percentages over all item categories ($\frac{\sum boot_j}{8}$).

2.5 | Results

Target trials, trials with response errors, with unusually slow (i.e., 1,500 ms or more) or unusually fast (i.e., 150 ms or faster) response times were excluded from the analysis. 1.58% of irrelevant and probe trials were excluded.

2.5.1 | Post-CIT recognition

For the final sample of 60 participants, target recognition accuracy after the RT-CIT was 95.2%. The probe recognition accuracy of the 30 guilty participants was 84.2%. Note that participants with more than two errors in either test were excluded (and replaced by another participant) and therefore not included in the final sample.

2.5.2 | Known-solution group analysis

Before we tried the searching CIT algorithms, we verified that participants in the guilty condition showed larger RT-CIT effects than participants in the innocent condition. A one-sided Bayesian independent samples *t* test on the CIT-effect between the guilty and the innocent group (using a weakly informative Cauchy prior; scale = 0.707) was used to compare the hypothesis H_1 (larger CIT-effects for the guilty group compared to the innocent group) to H_0 (no difference in the CIT-effect between the two groups). The test revealed very strong evidence for H_1 ($BF_{10} = 2.82^{*}10^{6}$) with a between-group effect size $dCIT_{between} = 1.76$ (95% credible interval [1.09, 2.28]) showing that the guilty group ($M\ dCIT_{within} = 0.36$; $SD = 0.22$) has larger within-participant CIT-effects than the innocent group ($M\ dCIT_{within} = 0.04$; $SD = 0.14$; see Table 2).[4]

2.5.3 | Known solution participant classification

We plotted the receiver operating characteristics (ROC) curve to assess if $dCIT_{within}$ can be used to discriminate between guilty and innocent participants. The area under the ROC curve (AUC) is an often-used index of diagnostic power (Fawcett, 2006). The AUC was 0.91 (95% CI: [0.84, 0.98]) and well above chance level. We used leave-one-out cross-validation (LOO CV) and applied the cut-off that maximizes the Youden's *J* statistic (*J* = sensitivity + specificity −1; Youden, 1950) in the model building sample for individual classification. This procedure achieved a cross-validated classification accuracy of 85% (25 of 30 or 83% of guilty participants and 26 of 30 or 87% of innocent participants were classified correctly). For sake of comparison, we note that the commonly used cutoff $d = 0.2$ (see Noordraven & Verschuere, 2013) led to a specificity of 86.7% and a sensitivity of 80% (overall accuracy 83.3%).

The reason Youden's J was used is that it is not biased by the base rate of guilty people in the population under investigation because it weights an increase of 1% in sensitivity and a 1% increase in specificity equally. To illustrate this, let us assume 10,000 people with a guilty base rate of 1% (i.e., 100 guilty, 9,900 innocent people) should be classified. A 5% increase in sensitivity will classify five additional guilty people as such which is an increase in accuracy of 0.05% whereas a 5% increase in specificity will classify 495 additional innocent people correctly which is an increase of 4.95%. However, Youden's J will in both cases increase by 0.05. Why this is a desirable

TABLE 2 Mean reaction times (in ms; *SD*s in parentheses), CIT-effect of innocent and guilty participants by item type and enactment (enacted, intention, collapsed), and between group effect sizes by enactment

	Innocent				Guilty				
	RT				RT				
	Irrelevant	Probe	Target	$dCIT_{within}$	Irrelevant	Probe	Target	$dCIT_{within}$	$dCIT_{between}$
Enacted	482 (154)	481 (152)	600 (132)	−0.00 (0.19)	486 (146)	532 (181)	627 (135)	0.36 (0.30)	1.45
Intention	491 (151)	501 (162)	618 (139)	0.08 (0.21)	502 (151)	551 (182)	646 (148)	0.39 (0.34)	1.07
Collapsed	486 (153)	491 (158)	609 (136)	0.04 (0.14)	494 (148)	541 (182)	637 (142)	0.36 (0.22)	1.76

property becomes evident when comparing classifiers from different scenarios: A naïve classifier that classifies everyone as innocent would reach 99% accuracy (but $J = 0$) in this example—an almost perfect classifier with 100% true positive and 2% false positive would reach 98% accuracy with $J = 0.98$. The same two classifiers with a guilty base rate of 50% would achieve accuracies of 50% and 99% while J remains unchanged.

2.5.4 | Task enactment: Past versus future behaviour

A two-tailed Bayesian paired samples *t* test with a weakly informative Cauchy prior (scale $r = 0.707$) was conducted to compare the CIT-effects of guilty participants in enacted and intent items using JASP (JASP Team, 2019). We found moderate evidence ($B_{01} = 4.91$; 95%-credible interval [−0.52, 0.39]) that the null-hypothesis stating that the CIT-effects between enacted and intent items do not differ is better supported by the data than the alternative hypothesis that CIT-effects do differ.

2.5.5 | Searching CIT

Item classification performance of the searching algorithms was assessed by the Youden's J at the optimal cut-off. As explained above, accuracy is not a suitable measure when the base rates are very different from 0.5. A naïve classifier (e.g., "all items are irrelevant") would achieve 90% accuracy because 90% of the to-be-classified items belong to the irrelevant category. Also, the AUC used in the known solution participant classification cannot be used with the first to second bootstrap algorithm because only one item in each category is classified based on a criterion. This means that if at least one irrelevant item shows a larger mean RT than the probe in this category (the probe is therefore automatically classified as irrelevant), the algorithm will never reach a sensitivity of 1 no matter how liberal the criterion is set; which is a prerequisite to interpret the AUC.

Using LOO CV procedure for item classification, the standardization algorithm achieved a Youden's J of 0.37 (sensitivity = 0.68; specificity = 0.69). Participant classification was above chance with AUC = 0.68 (95% CI: [0.54, 0.82]) but significantly worse than the known solution CIT (DeLong's test for two ROC curves: $D = -2.98$; $p < .01$; Robin et al., 2011). LOO CV resulted in a classification accuracy of 65% (19 of 30 guilty participants and 20 of 30 innocent participants were classified correctly).

The first to second bootstrap algorithm for item classification achieved a cross-validated Youden's J of 0.33 (sensitivity = 0.50; specificity = 0.83). It should be noted that this algorithm cannot achieve any arbitrary sensitivity or specificity; it strongly depends on the number of the probes that are selected as possible probes in the first step of the algorithm (in the current study, 120 of 240 probes). Participant classification was above chance level with an AUC of 0.74 (95% CI: [0.61, 0.87]) but significantly worse than the known solution CIT (DeLong's test for two ROC curves: D = −2.33; $p = 0.02$; Robin et al., 2011). LOO CV resulted in a classification accuracy of 68% (18 of 30 guilty participants and 23 of 30 innocent participants were classified correctly).

2.6 | Discussion

In Study 1, we conducted a CIT in an airport setting. The known solution CIT—to test crime knowledge and to investigate if items related to an enacted task show larger CIT-effects than items related to intentions—showed larger CIT-effects for guilty than for innocent participants with a classification performance (85% accuracy) well above chance. We found no evidence for an effect of enactment (i.e., differences in the CIT-effect for items related to enacted tasks and items related to future intentions). Although we tried to make the enacted and intent items and categories comparable (e.g., each contained one alphanumeric non-word and neither contained emotionally loaded items), the mock crime scenario did not allow for counterbalancing between enacted and intent items. The possibility that this null effect is due to item selection can therefore not be discarded completely.

For the first time, we showed that RTs can be used to find crime relevant information and distinguish knowledgeable from naïve participants using the searching RT-CIT. Both searching algorithms achieved item and participant classification on a similar and above chance level. While encouraging and providing initial evidence for the validity of the searching RT-CIT, and therefore new applications, it was also evident that the known solution CIT is substantially more accurate in classifying participants.

3 | STUDY 2: SIMULATION STUDY

The results of Study 1 warrant further exploration on how the searching algorithms are influenced by different factors. We ran a simulation study to do so. We simulated a wide array of datasets that varied along the dimensions of CIT-effect size (eight levels: $dCIT_{within}$ = 0.2, 0.3, 0.36, 0.4, 0.45, 0.5, 0.6, 1), number of item categories (i.e., information that could be tested; eight levels: 1–8), and number of trials per item (five levels: 5, 10, 20, 50, 100). This resulted in a total of 320 (8*8*5) datasets with 1,000 simulated participants each (500 guilty, 500 innocent). The effect sizes $dCIT_{within}$ = 0.36 and $dCIT_{within}$ = 0.45 were simulated to compare the algorithms' performance on the simulated data to their performance on empirical data with the same effect sizes. The effect size of $dCIT_{within}$ = 0.36 was the CIT-effect found in Study 1, $dCIT_{within}$ = 0.46[5] corresponds to the CIT-effect found in an independent study (Verschuere & Kleinberg, 2016) whose data will be used to validate the simulations in Study 3.

3.1 | Data generation

Data were generated on the trial level with the following assumptions: (a) People differ in their baseline reaction times. As an estimate of baseline RTs, we used the reaction time of innocent participants to irrelevant items. (b) Knowledgeable participants differ in their response to the probe (e.g., due to different perceived salience of the probes, or different ability to suppress the initial YES response to probes) which results in different CIT-effects in knowledgeable participants. (c) Innocent participants do not recognize the probe and therefore do not show a CIT-effect. (d) Reaction times on every trial are influenced by unsystematic noise. For now, we did not include item-effects or effects of item category.

Following these assumptions, the response time for participant *i* on trial *j* is generated by adding the different components. For innocent participants and for irrelevant items of guilty participants this results in $RT_{i,j}$ = baseline RT_i + $noise_j$ and for probe trials of guilty participants it is $RT_{i,j}$ = baseline RT_i + CIT-$effect_i$ + $noise_j$. For both irrelevant and probe items, the noise was drawn from a right skewed distribution with a mean of 0 (exponentially modified Gaussian distributions with a mu = 0, sigma = 56, and beta = 120 for irrelevant items and mu = 0, sigma = 72, and beta = 154 for probes). These values were derived from fitting an exponentially modified Gaussian model to the data of Study 1 (see osf.io/69yrj/ for further information). Reaction times of targets were not simulated as they do not influence the analysis.

3.2 | Results

Figure 1 presents the searching CIT algorithms' performance on simulated data with the effect size found in Study 1 (d_{within} = 0.36) but without any item effects (e.g., word length, numbers versus words, salience). The simulations show a mean maximized Youden's J of 0.47 (sensitivity = 0.68, specificity = 0.79) for the standardization algorithm when each item is presented 20 times compared to the empirically observed J of 0.37 in Study 1. The first to second bootstrap algorithm achieved a mean maximized J of 0.45 compared to the empirically found J of 0.33 (sensitivity = 0.60, specificity = 0.85). This implies that with the same effect size and under optimal conditions (i.e., no item effects) the algorithms could perform better than what we found in Study 1.

At least four conclusions can be drawn from the simulation outcomes. First, as observed in Study 1, the simulations indicate that, at the maximized Youden's J, the standardization algorithm is more liberal (i.e., is more likely to categorize items as probes). This results in a somewhat higher sensitivity but also a somewhat lower specificity than the first to second bootstrap algorithm. Second, across all the simulations (see Appendix B) the standardization algorithm tends to outperform the first to second bootstrap algorithm. Third, evidently, the algorithms perform better when the $dCIT_{within}$ is larger (Figure 2). Fourth, the simulations indicate that both algorithms would profit from increasing the number of repetitions per item.

3.3 | Discussion

The goal of Study 2 was to explore the searching algorithms' performance in the absence of item effects under various conditions. Both algorithms show very similar benefits from more repetitions per item and from larger CIT-effects.

The main differences in item classification performance between the two algorithms root in the sensitivity limitations of the first to second bootstrap algorithm. Its sensitivity cannot exceed the proportion of probes that has been marked as possible probe in the first step of the algorithm, no matter how liberal the criterion in the second step.

Direct comparison between the empirical data and the simulated data with the same effect size showed that item effects seem to influence the searching algorithms negatively. The performances based on the simulated data should therefore be considered estimates of the theoretical ceiling performance.

4 | STUDY 3: VALIDATION ON ARCHIVAL DATA

The simulation study showed that both algorithms have the potential to perform better than what we found in Study 1 if the number of item repetitions is increased or if the CIT effect is larger. Although this has not been manipulated in the simulation, the comparison between simulated and empirical data suggests that reducing item effects could have a considerable impact on the searching CIT performance also. To validate the algorithms on a second dataset and to show that the

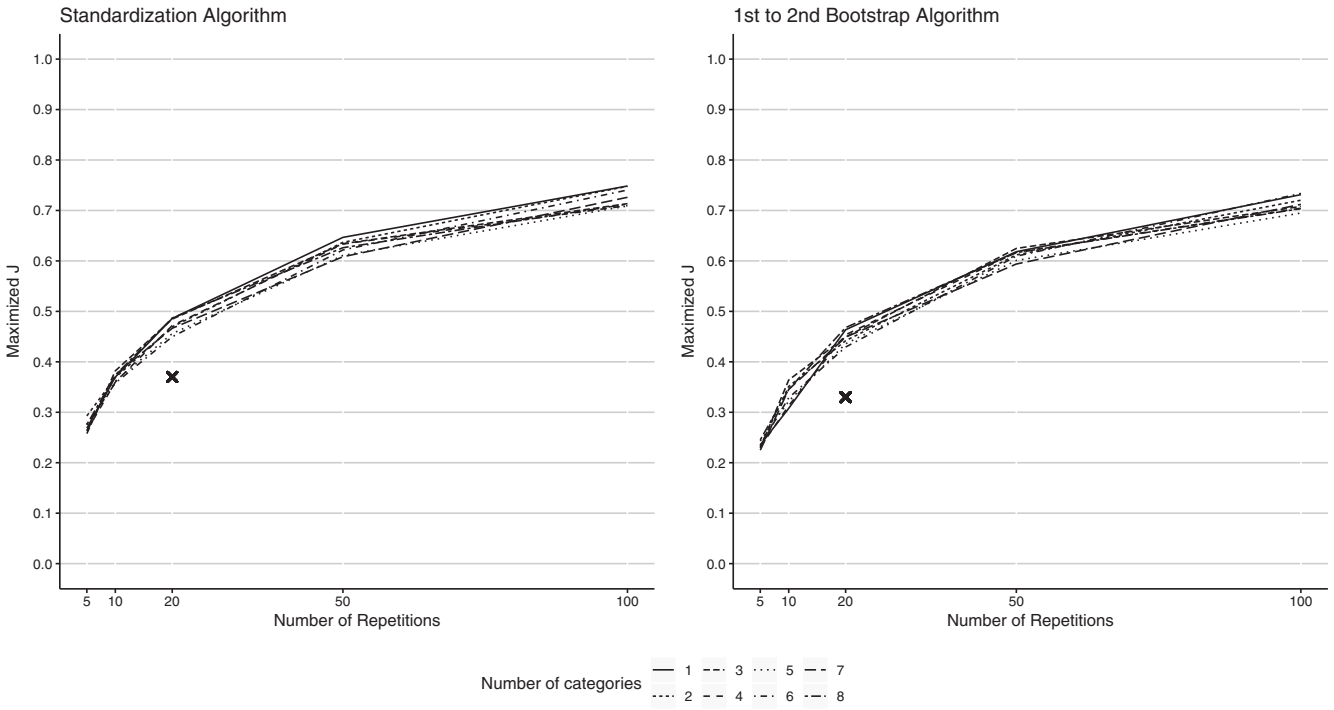

FIGURE 1 Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the first to second bootstrap algorithm (right) on simulated data with CIT-effect size $dCIT_{within}$ = 0.36 (lines) and on the empirical data of Study 1 (cross)

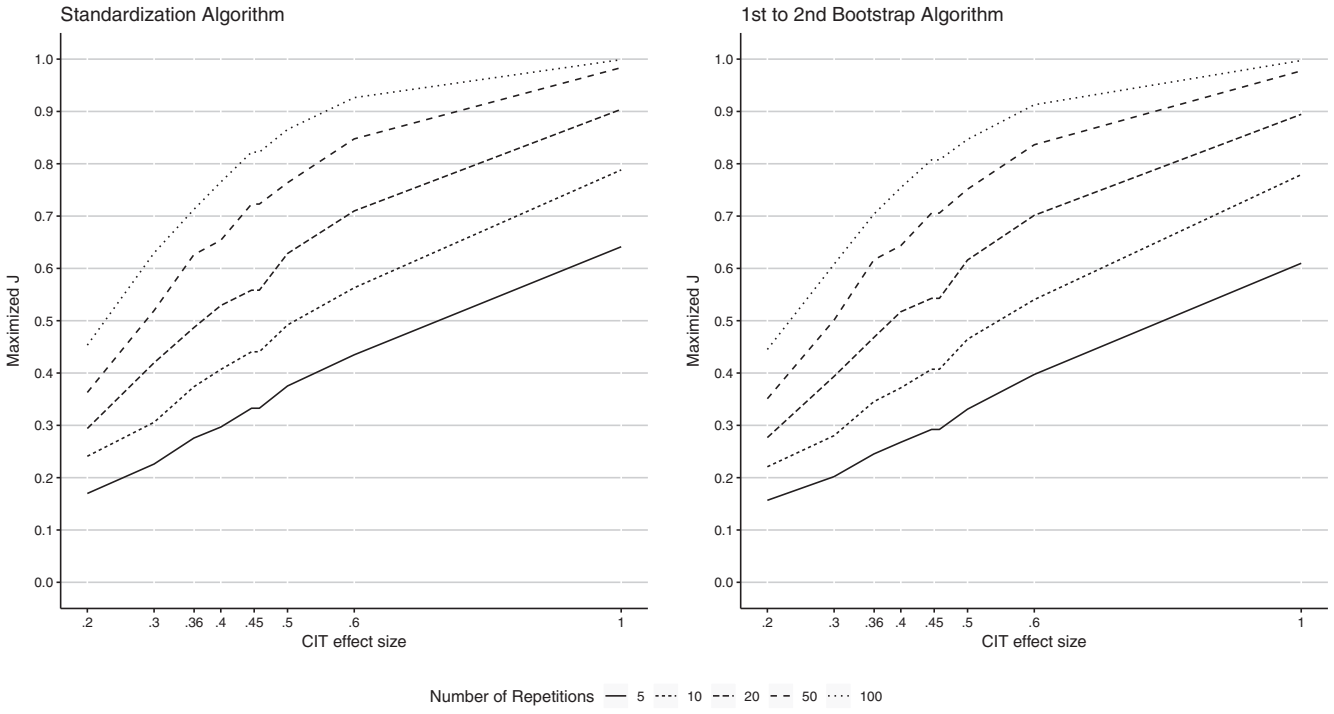

FIGURE 2 Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the first to second bootstrap algorithm (right) on the simulated dataset with eight item categories

simulations yield useful ceiling estimates, we applied the analysis from Study 1 on the autobiographical RT-CIT data from Verschuere and Kleinberg (2016). This study had the same number of repetitions as Study 1, larger CIT-effects (which is expected due to the high relevance of the autobiographical information), and possibly smaller items effect. Smaller item effects can be expected because there are no

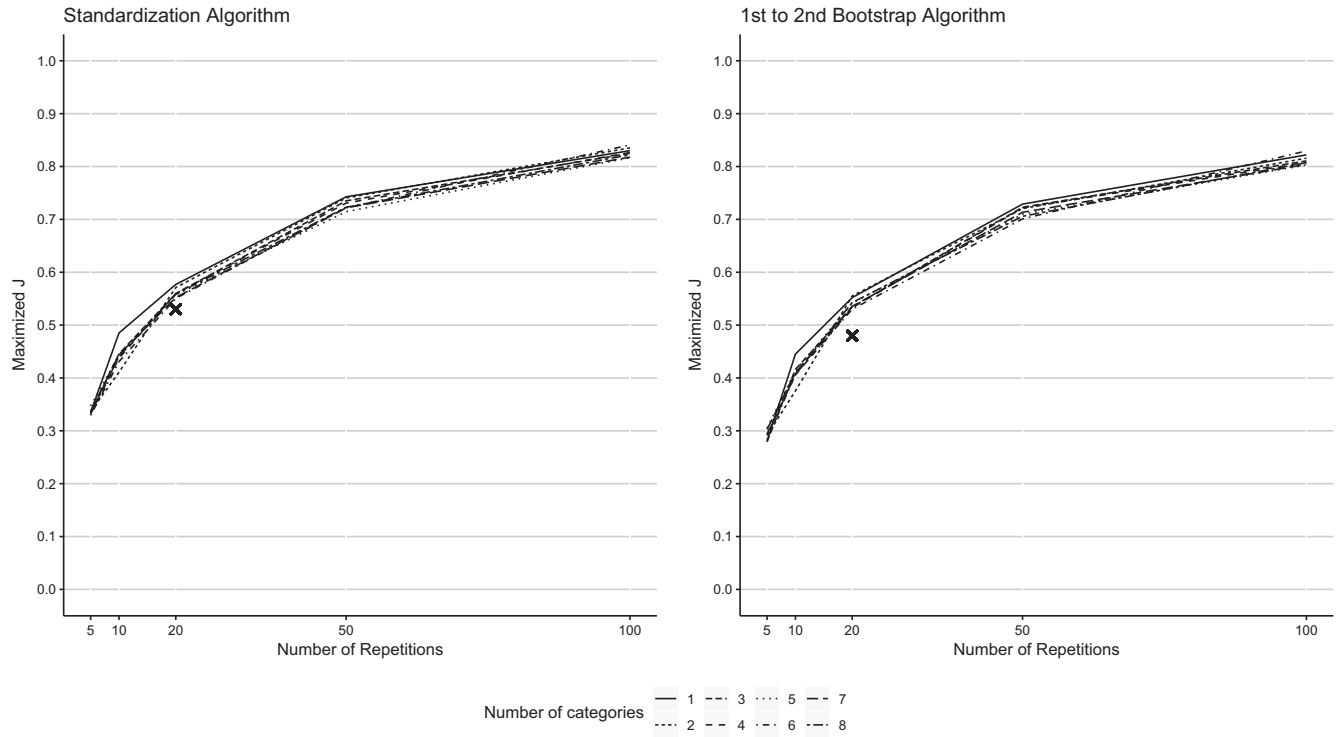

FIGURE 3 Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the first to second bootstrap algorithm (right) on simulated data with CIT-effect size $dCIT_{within}$ = 0.45 (lines) and on the empirical data of Study 3 (cross)

differences in how well the different information was learned (since the information is autobiographic and does not need to be learned) and because participants could indicate if an irrelevant item stood out to them which lead to the exclusion of that item, reducing saliency effects among the irrelevant items.

4.1 | Method

For a detailed description of the method, we refer the reader to Verschuere and Kleinberg (2016). The data can be found on osf.io/cg5es. In brief, an autobiographical RT-CIT was used with five item categories (first name, last name, university course, birthday, country of origin) and 20 trials per item for a total of 600 trials. Participants that were instructed to hide their identity showed a mean CIT-effect of $M(dCIT_{within})$ = 0.46 (SD = 0.23), whereas unknowledgeable participants had a CIT-effect of $M(dCIT_{within})$ = −0.01 (SD = 0.13).

4.2 | Results

Maximizing the Youden's J of the searching algorithms for the empirical data of Verschuere and Kleinberg (2016) using LOO CV resulted in a Youden's J of 0.53 (sensitivity = 0.83; specificity = 0.69) for the standardization and a J of 0.48 (sensitivity = 0.62; specificity = 0.86) for the first to second bootstrap algorithm. Therefore, both algorithms showed above-chance classification performance. We also obtained further indications that the standardization algorithm is more liberal (at the cost of lower specificity) and that it shows slightly better overall discriminability. Figure 3 visualizes the performance of the algorithms in comparison with the theoretical ceiling performance based on simulated data.[6]

Participant classification based on the searching algorithms was with AUC = 0.68 (95% CI: [0.56, 0.81]) for the standardization and AUC = 0.69 (95% CI: [0.56, 0.81]) for the first to second bootstrap algorithm above chance for both algorithms.

4.3 | Discussion

The aim of Study 3 was to validate the searching CIT algorithms on an independent dataset and to show that the simulations yield realistic results. As predicted by the simulations, both algorithms showed better item classification than in Study 1. Furthermore, the finding from Study 1 that the standardization algorithm is more liberal when the optimal cut-off is used was replicated.

Study 3, therefore, showed the validity of the searching algorithms on an independent dataset and presented additional evidence that our data simulation can be used to estimate the ceiling performance those algorithms can theoretically achieve given a certain effect size.

5 | GENERAL DISCUSSION

The present study used the RT-CIT in a mock-terror attack scenario at an international airport and explored the potential of two searching

algorithms to reveal critical information about the attack and to classify participants.

We first showed that the known solution RT-CIT can be applied in an airport setting with a high classification accuracy of 85% (AUC = 0.91; using the commonly used cutoff of $dCIT_{within}$ = .2 resulted in an accuracy of 83.3%). This shows that high accuracies can also be achieved in situations with possibly higher agitation levels (due to high security standards, police presence, and the unfamiliar airport environment) than studies conducted in university settings.

Especially in the terror context, the police are interested in detecting malicious intent to prevent an attack. To investigate if intentions can be detected to the same degree as past actions, we compared the CIT-effect of items that the participants physically interacted with to items related to their intentions. We found moderate evidence that the CIT-effect is not influenced by enactment. These results could be explained in different ways: It could be that the richness of the memory trace (if the memory of enacted and intent items is sufficiently strong) indeed does not influence the CIT-effect. An alternative explanation could be that the effect was masked by an increased focus on the intent items as they were still relevant to execute the mock crime successfully. Theoretically, because the mock crime scenario did not allow us to balance the items between the past action and intent condition, this finding could also be a result of the item selection. Although we cannot definitely conclude that enactment does not influence the CIT-effect, our results provide further evidence that the CIT is well suited to detect memory of past actions and intentions.

Finally, we set out to investigate whether response times can be used to reveal new crime details to the investigative party. Study 1 showed that searching CIT algorithms can be used to identify crime relevant information above chance level. The standardization algorithm showed slightly higher discriminability, but the main difference was that its sensitivity was higher at the cost of lower specificity compared to the first to second bootstrap algorithm. However, this only applies when the algorithms are evaluated at their maximized Youden's J. When the criterion in the standardization procedure is set to match a certain sensitivity or specificity of the bootstrap procedure, they both achieve the same performance. Furthermore, the searching CIT achieved above chance classification performance of participants into guilty/innocent but both algorithms were considerably worse than the known solution CIT. Gathering useful information before testing a suspect with the CIT is therefore still needed to get the most accurate guilty/innocent classification.

To explore the algorithms' potential under different conditions and without item effects, we turned to simulated data. Whereas our simulation study sheds light on what would happen with different numbers of trials and different effect sizes, it is limited in the number of factors it takes into account and currently disregards known moderators of the CIT effect (e.g., saliency, countermeasures; Suchotzki et al., 2017). Both algorithms show very similar benefits from more repetitions per item and from larger CIT-effects. The simulations further indicated that the searching RT-CIT could achieve substantially better classification performance given the right conditions (i.e., increased CIT-effects and more repetitions per item, see Figure 2). Note that possible effects of habituation and fatigue could not be taken into account due to the lack of research in this area of the RT-CIT. Optimization of the paradigm to increase the CIT-effect is very challenging and will take time, but testing the validity of the RT-CIT with large numbers of trials and investigating the effects of fatigue and habituation might give valuable insight and could be done quickly. In addition, this knowledge could be used to refine the simulations. Especially in the exploration phase of this new field of searching algorithms in the RT-CIT, data simulation could be a valuable tool to explore the properties of algorithms. Using simulated data to explore the behaviour of a system (e.g., algorithm, computational model) in a wide array of conditions is well established in cognitive psychology (Sun, 2008) and could be a promising direction for CIT research in general.

The validation of the results from our simulations using independent data is further evidence that our data simulation can be used to estimate the maximal performance those algorithms can theoretically achieve given a certain effect size. The remaining discrepancy between the performance on the simulated and empirical data is most likely due to item effects in dimensions that are likely to influence the CIT-effect or response times in general, such as saliency (Kleinberg & Verschuere, 2015; Verschuere, Kleinberg, & Theocharidou, 2015), word length (Barton, Hanif, Eklinder Björnström, & Hills, 2014), and word frequency (Rayner & Duffy, 1986).

In general, both algorithms show very similar classification performance with a slight advantage for the standardization algorithm, especially with small numbers of repetitions per item. Possibly the most relevant difference between the algorithms is that the classification criteria of the standardization algorithm can be set freely to achieve any desired sensitivity/specificity, whereas the first to second bootstrap algorithm's sensitivity is limited to the proportion of probes that are considered “possible probe” in the first step. How the criterion should be set in practice is determined by the circumstances. If high sensitivity is needed (e.g., terror prevention) the criterion is set lower than in scenarios with very limited resources that must not be spent on false alarms—a flexibility that is not achieved by the first to second bootstrap algorithm.

Another important difference between the algorithms is the susceptibility to an irrelevant item showing a CIT-effect. This could be due to an involuntary reaction of a participant to an item or it could be part of a countermeasure used by guilty participants. While both algorithms are expected to be affected in a similar way for innocent participants, the effect for guilty participants can be different. Let us assume a CIT-effect of that irrelevant item is of the same size as the actual probe. For the standardization algorithm, this would result in the same $dCIT_{i,j}$ score for this irrelevant item as for the probe, but they would still be larger than zero and therefore diagnostic. Note that the $dCIT_{i,j}$ score of the probe would be smaller than without an irrelevant signal because the difference of the means decreases and the *SD* of irrelevants increases (see Study 1 for the formula). Using the optimal cut-off, both items would be classified as probes while the other irrelevant items of guilty and innocent participants would still be

TABLE 3 Overview of the searching algorithms' evaluation

	Standardization algorithm	First to second bootstrap algorithm
Classification performance	Fair	Fair
Sensitivity space	Unrestricted	Restricted
Vulnerability to countermeasures	Vulnerable	Very vulnerable
Computational resources needed	Low	Medium

classified correctly. Furthermore, $M(\max[dCIT_{i,j}])$ could still be used to classify participants. The first to second bootstrap algorithm would, in the first step, only treat half of the real probes as "possible probes" limiting the sensitivity to a maximum of 0.5. The bootstrap comparison in step two would take place between the real probe and the irrelevant item that showed the same CIT-effect, yielding a mean of 50%—the same as when two irrelevant items of an innocent participant are compared. In this case, the second step will not improve the classification. The best performance would be reached when every possible probe is classified as the probe. Participant classification, however, would not be possible since this guilty participant would show no bootstrap difference, just like innocent participants.

The possibility of stronger countermeasures that result in a CIT-effect of the irrelevant item larger than the CIT-effect of the probe or of applying countermeasures to multiple irrelevant items must be considered also. These possibilities lead to a further decrease in classification performance for both algorithms but to a lesser extent in the standardization than in the first to second bootstrap algorithm, following the same rational as in the presented example.

From a practical point of view, the standardization algorithm has the advantage that it takes less computational resources and therefore less time, since it does not rely on bootstrapping. For those reasons, we conclude that the standardization algorithm has more desirable properties and should currently be favoured over the first to second bootstrap algorithm (Table 3).

Finally, a general limitation of the searching CIT needs to be considered. In this study, as in most others, the true probe was always present in the searching CIT which does not need to be the case in practice. Real-life situations rarely have a closed set possible probes in which the investigator knows that the real probe is included. This either means that the searching CIT (irrespective of the measures used) should only be applied in very rare situations or that an item that covers all other possibilities should be included. The latter is done in Japan, the only country that uses the searching CIT on a large scale (Osugi, 2018). The effects of this practice have yet to be thoroughly investigated.

5.1 | Applicability

The results from the empirical data and the simulations suggest that the RT-CIT is suitable not only to test if someone possesses specific crime knowledge (known solution CIT) but also to find unknown crime information among plausible but crime unrelated alternatives (searching CIT). The known solution RT-CIT could already be applied in specific situations such as testing a suspect for crime knowledge as part of a police investigation, at the border when the police suspect that the country of origin provided by a person is wrong and they have a specific suspicion where the person might be from (e.g., by testing knowledge about lesser known towns of that country), or possibly testing the knowledge about a substance found in a passengers' luggage which the passenger claims not to have packed. Our finding suggests, that it can also be used to test for intentions such as plans for a journey or a terror attack.

The searching algorithms open up an additional spectrum of scenarios in which the RT-CIT can be applied. In the context of an investigation, for example when the police caught someone carrying illegal substances, they could use the searching RT-CIT to narrow down or prioritize where to look for the seller; or in a situation where the police have some information about a planned terror attack but do not know where it will take place, but they have a suspect that they believe to have knowledge about the attack. It could be used to get hints on where the attack will take place, what kind of bomb to look for, the day of the attack and alike. Although not addressed in this study, the searching RT-CITs performance can most likely be increased if multiple people sharing the same crime knowledge can be tested, as it has been done with the physiological CIT (e.g., Breska et al., 2014; Breska, Ben-Shakhar, & Gronau, 2012; Elaad, 2016). This reduces the impact of one person showing a distinct reaction to an irrelevant item for any reason.

5.2 | Limitations

Although we used a highly realistic scenario in Study 1 by getting participants to the airport, making them execute the mock attack in a high security environment with real police present, and a believable cover story, three important aspects are very different from a real-life scenario. (a) Apart from not getting the monetary bonus of 5 CHF, there were no negative consequences being classified as guilty. In reality, this would be an extremely high stakes crime and the suspects would be very motivated not to be classified as guilty. Although high stakes crimes need more investigation, the meta-analysis of Suchotzki et al. (2017) did not find an effect of motivation on the size of the CIT-effect. (b) The participants were given instructions about the mock crime, learned them and then planned the execution. Planning the attack from scratch and considering possible alternatives might impact the CIT-effect of alternatives that were considered but not chosen, which might influence the classification performance of both the known solution and the searching CIT. (c) We used a student population, which is not representative of the general population.

The sample size of $N = 60$ of Study 1 was not enough to find conclusive evidence for the null hypothesis that stated that there is no effect of enactment, even though the difference in the CIT-effect was minimal. Although the results are promising, studies with larger sample sizes are needed to reach conclusive evidence on the matter.

As with any deception detection tool that might get used in practice, its susceptibility to countermeasures is an important concern that needs to be addressed. Suchotzki et al. (2017) found RT-based deception detection measures to be vulnerable to countermeasures but further research is warranted as different RT-based paradigms had to be analyzed together due to the few countermeasure studies that were conducted. Furthermore, susceptibility to countermeasures does not necessarily mean that detection methods cannot be used. If countermeasures can be detected, this can also be a valuable piece of information to the examiner by itself.

5.3 | Future studies

Searching algorithms on RT-CIT data is a research field that remains to be explored. We encourage other researchers to develop new algorithms and use simulated data to explore boundary conditions. Promising research directions include non-binary classifications (i.e., providing a measure of certainty that the classification is correct), machine learning approaches, and using converging evidence of multiple algorithms.

ACKNOWLEDGMENTS

We thank the Swiss Federal Office of Civil Aviation for the financial support (project number: 2016-106). We thank the Zurich State Police, Airport Division for their financial support and the possibility to use their infrastructure to conduct Study 1. We thank Zoé Dolder for her help in data collection.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data and R-scripts that support the findings of this study are openly available on OSF https://osf.io/69yrj/.

ORCID

Dave Koller https://orcid.org/0000-0001-5315-8782

Bruno Verschuere https://orcid.org/0000-0002-6161-4415

ENDNOTES

[1] AUC = 0.5 represents chance performance; AUC = 1 is perfect classification performance.

[2] The study was preregistered (osf.io/69yrj/) but that preregistration was premature and the authors decided to analyse the classification based on adaptations of already existing algorithms and refrained from calculating the preregistered Bayesian index I. Since this is an integral part of the study, it should be considered exploratory.

[3] The innocent group did not complete this recognition test because the probes and targets were counterbalanced. However, innocent participants were presented with all the items and were asked to indicate the most plausible one for each category. χ^2-tests for each item category showed no significant effects after correcting for multiple comparisons (Bonferroni).

[4] CIT-effects were calculated as: $dCIT_{between} = \frac{M(dCITguilty) - M(dCITinnocent)}{\sqrt{var(dCITguilty) + var(dCITinnocent)/2}}$; $dCIT_{within} = \frac{M(RTprobe) - M(RTirrelevant)}{SD(RTirrelevant)}$.

[5] $dCIT_{within}$ = 0.45 is based on initial calculations that contained a mistake in data aggregation. The correct effect size is $dCIT_{within}$ = 0.46 but we refrained from running new simulations because the difference is minimal.

[6] Applying the searching algorithms on simulated data with M ($dCIT_{within}$) = 0.45 resulted in J = 0.56 (sensitivity = 0.74; specificity = 0.81) and J = 0.54 (sensitivity = 0.68; specificity = 0.86) for the standardization and first to second bootstrap algorithm respectively.

REFERENCES

Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology*, *31* (5–6), 378–412. https://doi.org/10.1080/02643294.2014.895314

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Breska, A., Ben-Shakhar, G., & Gronau, N. (2012). Algorithms for detecting concealed knowledge among groups when the critical information is unavailable. *Journal of Experimental Psychology: Applied*, *18*(3), 292–300. https://doi.org/10.1037/a0028798

Breska, A., Zaidenberg, D., Gronau, N., & Ben-Shakhar, G. (2014). Psychophysiological detection of concealed information shared by groups: An empirical study of the searching CIT. *Journal of Experimental Psychology: Applied*, *20*(2), 136–146. https://doi.org/10.1037/xap0000015

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, *22*(1), 267–281. https://doi.org/10.1111/j.1467-9450.1981.tb00402.x

Elaad, E. (2016). Extracting critical information from group members' partial knowledge using the searching concealed information test. *Journal of Experimental Psychology: Applied*, *22*(4), 500–509. https://doi.org/10.1037/xap0000101

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

JASP Team. (2019). JASP (Version 0.9.1.0) [Computer software].

Klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2017). Unraveling the roles of orienting and inhibition in the concealed information test. *Psychophysiology*, *54*(4), 628–639. https://doi.org/10.1111/psyp.12825

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test memory. *PLoS ONE*, *10*(4), 1–17. https://doi.org/10.1371/journal.pone.0118715

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385–388. https://doi.org/10.1037/h0046060

Meijer, E. H., Bente, G., Ben-Shakhar, G., & Schumacher, A. (2013). Detecting concealed information from groups using a dynamic questioning approach: Simultaneous skin conductance measurement and immediate feedback. *Frontiers in Psychology*, *4*, 1–6. https://doi.org/10.3389/fpsyg.2013.00068

Meijer, E. H., Klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the concealed information test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, *51*(9), 879–904. https://doi.org/10.1111/psyp.12239

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*, *53*(5), 593–604. https://doi.org/10.1111/psyp.12609

Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology*, *48*(2), 149–154. https://doi.org/10.1111/j.1469-8986.2010.01050.x

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based concealed information test. *Applied Cognitive Psychology*, *27*(3), 328–335. https://doi.org/10.1002/acp.2910

Osugi, A. (2011). Daily application of the concealed information test: Japan. Verschuere B., Ben-Shakhar G., Meijer E. (Eds.), In *Memory*

detection: Theory and application of the concealed information test (pp. 253–275). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511975196.015

Osugi, A. (2018). Field findings from the concealed information test in Japan. In J. P. Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 97–121). Cambridge: Academic Press. https://doi.org/10.1016/C2016-0-03911-6

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition*, *14*(3), 191–201. https://doi.org/10.3758/BF03197692

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess 'guilty knowledge'. *Journal of Applied Psychology*, *85*(1), 30–37. https://doi.org/10.1037/0021-9010.85.1.30

Suchotzki, K., Verschuere, B., van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/bul0000087

Sun, R. (2008). In R. Sun (Ed.), *The Cambridge handbook of computational psychology*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511816772

The MathWorks. (2018). *MATLAB*. Natick, MA.

Verschuere, B., & Kleinberg, B. (2016). ID-check: Online concealed information test reveals true identity. *Journal of Forensic Sciences*, *61*, S237–S240. https://doi.org/10.1111/1556-4029.12960

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition*, *4*(1), 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Koller D, Hofer F, Grolig T, Ghelfi S, Verschuere B. What are you hiding? Initial validation of the reaction time-based searching concealed information test. *Appl Cognit Psychol*. 2020;1–13. https://doi.org/10.1002/acp.3717